# STRATIFIED SAMPLING

Jeff Wooldridge
Michigan State University
LABOUR Lectures, EIEF
October 18-19, 2011

## 1. The Basic Methodology

- With stratified sampling, some segments of the population are overrepresented or underrepresented by the sampling scheme. If we know enough information about the stratification scheme (and the population), we can modify standard econometric methods and consistently estimate population parameters.
- There are two common types of stratified sampling, standard stratified (SS) sampling and variable probability (VP) sampling. A third type of sampling, typically called multinomial sampling, is practically indistinguishable from SS sampling, but it generates a random sample from a modified population.

- **SS Sampling**: Partition the sample space, say $\mathcal{W}$, into $G$ non-overlapping, exhaustive groups, $\{\mathcal{W}_g : g = 1, \ldots G\}$. A random sample is taken from each group $g$, say $\{w_{gi} : i = 1, \ldots, N_g\}$, where $N_g$ is the number of observations drawn from stratum $g$ and $N = N_1 + N_2 + \ldots + N_G$ is the total number of observations.
- Let $w$ be a random vector representing the population. Each each random draw from stratum $g$ has the same distribution as $w$ conditional on $w$ belonging to $\mathcal{W}_g$:

$$D(w_{gi}) = D(w|w \in \mathcal{W}_g),\ i = 1, \ldots, N_g. \tag{1.1}$$

- We only know we have an SS sample if we are told.

• What if we want to estimate the mean of $w$ from an SS sample? Let $\pi_g = P(w \in \mathcal{W}_g)$ be the probability that $w$ falls into stratum $g$; the $\pi_g$, which are population frequencies, are often called the "aggregate shares." If we know the $\pi_g$ (or can consistently estimate them), then $\mu_w = E(w)$ is identified by a weighted average of the expected values for the strata:

$$\mu_w = \pi_1 E(w|w \in \mathcal{W}_1) + \ldots + \pi_G E(w|w \in \mathcal{W}_G). \tag{1.2}$$

• Sometimes the $\pi_g$ are obtained from census data.

- An unbiased estimator of $\mu_w$ is obtained by replacing each $E(w|w \in \mathcal{W}_g)$ with its unbiased estimator, the sample average from stratum $g$:

$$\hat{\mu}_w = \pi_1\bar{w}_1 + \pi_2\bar{w}_2 \ldots + \pi_G\bar{w}_G, \tag{1.3}$$

where $\bar{w}_g$ is the sample average from stratum $g$.

• As the strata sample sizes grow, $\hat{\mu}_w$ is also a consistent estimator of $\mu_w$. It is sufficient to assume $N_g/N \to \eta_g > 0$ for $g = 1, \ldots, G$.

• The variance is easy to calculate because the sample averages are independent across strata and the sampling is random within each stratum:

$$\begin{aligned} Var(\hat{\mu}_w) &= \pi_1^2 Var(\bar{w}_1) + \ldots + \pi_G^2 Var(\bar{w}_G) \\ &= \pi_1^2(\sigma_1^2/N_1) + \ldots + \pi_G^2(\sigma_G^2/N_G) \end{aligned} \tag{1.4}$$

• Each $\sigma_g^2$ can be estimated using the usual unbiased variance estimator:

$$\hat{\sigma}_g^2 = (N_g - 1)^{-1} \sum_{i=1}^{N_g} (w_{gi} - \bar{w}_g)^2 \tag{1.5}$$

Thus,

$$\widehat{Var}(\hat{\mu}_w) = \pi_1^2(\hat{\sigma}_1^2/N_1) + \ldots + \pi_G^2(\hat{\sigma}_G^2/N_G)$$

and so the standard error of $\hat{\mu}_w$ is

$$se(\hat{\mu}_w) = [\pi_1^2(\hat{\sigma}_1^2/N_1) + \ldots + \pi_G^2(\hat{\sigma}_G^2/N_G)]^{1/2}. \tag{1.6}$$

- Useful to have a formula for $\hat{\mu}_w$ as a weighted average across all observations:

$$\begin{aligned}\hat{\mu}_w &= (\pi_1/h_1)N^{-1}\sum_{i=1}^{N_1} w_{1i} + \ldots + (\pi_G/h_G)N^{-1}\sum_{i=1}^{N_G} w_{Gi} \\ &= N^{-1}\sum_{i=1}^{N}(\pi_{g_i}/h_{g_i})w_i \end{aligned} \tag{1.7}$$

where $h_g = N_g/N$ is the fraction of observations in stratum $g$ and in (1.7) we drop the stratum index on the observations.

• **Variable Probability Sampling**: Often used where little, if anything, is known about respondents ahead of time. Still partition the sample space, but a unit is drawn at random from the population. If the observation falls into stratum $g$, it is kept with (nonzero) sampling probability, $p_g$. That is, random draw $w_i$ is kept with probability $p_g$ if $w_i \in \mathcal{W}_g$.

• The population is sampled $N$ times ($N$ not always reported with VP samples). We always know how many data points were kept; call this $M$ – a random variable. Let $s_i$ be a selection indicator, equal to one if observation $i$ is kept. So $M = \sum_{i=1}^{N} s_i$.

- Let $\mathbf{z}_i$ be a $G$-vector of stratum indicators for draw $i$, that is, $z_{ig} = 1$ if and only if $w_i \in \mathcal{W}_g$. Because each draw is in one and only one stratum, $z_{i1} + z_{i2} + \ldots + z_{iG} = 1$.
- We can define

$$p(\mathbf{z}_i) = p_1 z_{i1} + \ldots + p_G z_{iG} \tag{1.8}$$

as the function that delivers the sampling probability for any random draw $i$.

- Key assumption for VP sampling: Conditional on being in stratum $g$, the chance of keeping an observation is $p_g$.
- Statistically, conditional on $\mathbf{z}_i$ (knowing the stratum), $s_i$ and $w_i$ are independent:

$$P(s_i = 1|\mathbf{z}_i, w_i) = P(s_i = 1|\mathbf{z}_i) \tag{1.9}$$

- Condition (1.9) means that the missingness of the data is *ignorable*. It is easy to show that

$$E[(s_i/p(\mathbf{z}_i))w_i] = E(w_i). \tag{1.10}$$

- Equation (1.10) is the key result for VP sampling. It says that weighting a selected observation by the inverse of its sampling probability allows us to recover the population mean. It is a special case of IPW estimation for general missing data.
- It follows that

$$N^{-1}\sum_{i=1}^{N}(s_i/p(\mathbf{z}_i))w_i \tag{1.11}$$

is a consistent estimator of $E(w_i)$. We can also write (1.11) as

$$(M/N)M^{-1}\sum_{i=1}^{N}(s_i/p(\mathbf{z}_i))w_i. \tag{1.12}$$

If we define weights as $\hat{v}_i = \hat{\rho}/p(\mathbf{z}_i)$ where $\hat{\rho} = M/N$ is the fraction of observations retained from the sampling scheme, then (1.12) is

$$M^{-1}\sum_{i=1}^{M}\hat{v}_i w_i, \tag{1.13}$$

where only the observed points are included in the sum.

• So, can write the estimator as a weighted average of the observed data points. If $p_g < \hat{\rho}$, the observations for stratum $g$ are underpresented in the eventual sample (asymptotically), and they receive weight greater than one.

## 2. Linear Regression Analysis

- Almost any estimation method can be used with SS or VP sampled data: OLS, IV, MLE, quasi-MLE, nonlinear least squares, quantile regression.

- Linear population model:

$$y = \mathbf{x}\boldsymbol{\beta} + u. \tag{2.1}$$

Two assumptions on $u$ are

$$E(u|\mathbf{x}) = 0 \tag{2.2}$$

$$E(\mathbf{x}'u) = \mathbf{0}. \tag{2.3}$$

• $E(\mathbf{x}'u) = \mathbf{0}$ is enough for consistency, but $E(u|\mathbf{x}) = 0$ has important implications for whether or not to weight under exogenous sampling.

• SS Sampling: A consistent estimator $\hat{\boldsymbol{\beta}}$ is obtained from the "weighted" least squares problem

$$\min_{\mathbf{b}} \sum_{i=1}^{N} v_i \cdot (y_i - \mathbf{x}_i\mathbf{b})^2, \tag{2.4}$$

where $v_i = \pi_{g_i}/h_{g_i}$ is the weight for observation $i$. (Remember, the weighting used here is not to solve any heteroskedasticity problem; it is to reweight the sample in order to consistently estimate the population parameter $\boldsymbol{\beta}$.)

• Key Question: How can we conduct valid inference using $\hat{\boldsymbol{\beta}}$? One possibility: use the White (1980) "heteroskedasticity-robust" sandwich estimator. When is this estimator the correct one? If two conditions hold: (i) $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, so that we are actually estimating a conditional mean; and (ii) the strata are determined by the explanatory variables, $\mathbf{x}$.

• When the White estimator is not consistent, it is conservative.

• Correct asymptotic variance requires more detailed formulation of the estimation problem:

$$\min_{\mathbf{b}} \left\{ \sum_{g=1}^{G} \pi_g \left[ N_g^{-1} \sum_{i=1}^{N_g} (y_{gi} - \mathbf{x}_{gi}\mathbf{b})^2 \right] \right\}. \tag{2.5}$$

- Asymptotic variance estimator:

$$\widehat{Avar(\hat{\boldsymbol{\beta}})} = \left[ \sum_{i=1}^{N} (\pi_{g_i}/h_{g_i}) \mathbf{x}_i' \mathbf{x}_i \right]^{-1} \cdot \left\{ \sum_{g=1}^{G} (\pi_g/h_g)^2 \left[ \sum_{i=1}^{N_g} (\mathbf{x}_{gi}' \hat{u}_{gi} - \overline{\mathbf{x}_g' \hat{u}_g})(\mathbf{x}_{gi}' \hat{u}_{gi} - \overline{\mathbf{x}_g' \hat{u}_g})' \right] \right\} \cdot \left[ \sum_{i=1}^{N} (\pi_{g_i}/h_{g_i}) \mathbf{x}_i' \mathbf{x}_i \right]^{-1}. \tag{2.6}$$

- Usual White estimator ignores the information on the strata of the observations, which is the same as dropping the within-stratum averages, $\overline{\mathbf{x}_g'\hat{u}_g}$. The estimate in (2.6) is always *smaller* than the usual White estimate.
- Econometrics packages, such as Stata, have survey sampling options that will compute (2.6) provided stratum membership is included along with the weights. If only the weights are provided, the larger asymptotic variance is computed.

- One case where there is no gain from subtracting within-strata means is when $E(u|\mathbf{x}) = 0$ and stratification is based on $\mathbf{x}$.
- If we add the homoskedasticity assumption $Var(u|\mathbf{x}) = \sigma^2$ with $E(u|\mathbf{x}) = 0$ and stratification is based on $\mathbf{x}$, the weighted estmator is less efficient than the unweighted estimator. (Both are consistent.)

- The debate about whether or not to weight centers on two facts: (i) The efficiency loss of weighting when the population model satisfies the classical linear model assumptions and stratification is exogenous. (ii) The failure of the unweighted estimator to consistently estimate $\boldsymbol{\beta}$ if we only assume

$$y = \mathbf{x}\boldsymbol{\beta} + u, E(\mathbf{x}'u) = \mathbf{0}, \tag{2.7}$$

even when stratification is based on $\mathbf{x}$. The weighted estimator consistently estimates $\boldsymbol{\beta}$ under (2.7).

• Analogous results hold for maximum likelihood, quasi-MLE, nonlinear least squares, instrumental variables. If one knows stratum identification along with the weights, the appropriate asymptotic variance matrix (which subtracts off within-stratum means of the score of the objective function) is smaller than the form derived by White (1982). For, say, MLE, if the density of $y$ given $\mathbf{x}$ is correctly specified, and stratification is based on $\mathbf{x}$, it is better not to weight. (But there are cases – including certain treatment effect estimators – where it is important to estimate the solution to a misspecified population problem.)

• Findings for SS sampling have analogs for VP sampling, and some additional results.

1. The Huber-White sandwich matrix applied to the weighted objective function (weighted by the $1/p_g$) *is* consistent when the *known* $p_g$ are used.

2. An asymptotically more efficient estimator is available when the retention frequencies, $\hat{p}_g = M_g/N_g$, are observed, where $M_g$ is the number of observed data points in stratum $g$ and $N_g$ is the number of times stratum $g$ was sampled. (Is $N_g$ known? It could be.)

The estimated asymptotic variance in that case is

$$\widehat{Avar(\hat{\boldsymbol{\beta}})} = \left[\sum_{i=1}^{M} \mathbf{x}_i' \mathbf{x}_i / \hat{p}_{g_i}\right]^{-1} \cdot \left\{\sum_{g=1}^{G} \hat{p}_g^{-2} \left[\sum_{i=1}^{M_g} (\mathbf{x}_{gi}' \hat{u}_{gi} - \overline{\mathbf{x}_g' \hat{u}_g})(\mathbf{x}_{gi}' \hat{u}_{gi} - \overline{\mathbf{x}_g' \hat{u}_g})'\right]\right\} \cdot \left[\sum_{i=1}^{M} \mathbf{x}_i' \mathbf{x}_i / \hat{p}_{g_i}\right]^{-1}, \tag{2.8}$$

where $M_g$ is the number of observed data points in stratum $g$.

Essentially the same as SS case in (2.6).

• If we use the known sampling weights, we drop $\overline{\mathbf{x}_g'\hat{u}_g}$ from (2.8). If $E(u|\mathbf{x}) = 0$ and the sampling is exogenous, we also drop this term because $E(\mathbf{x}'u|\mathbf{w} \in \mathcal{W}_g) = \mathbf{0}$ for all $g$, and this is whether or not we estimate the $p_g$.

• In Stata, use the "svyset" command, and then the "svy" prefix for sample statistics and econometric methods.

• Following example is with 6 strata and variable probability sampling in addition to different strata weights. (Within each stratum, VP sampling is used.)

```
. use http://www.stata-press.com/data/r10/nmihs


. des  idnum stratan finwgt marital age race birthwgt

              storage  display     value
variable name   type   format      label      variable label
-----------------------------------------------------------------------------
idnum           long   %10.0f                 ID number
stratan         byte   %8.0g                  Strata indicator 1-6
finwgt          double %10.0g                 Adjusted sampling weight
marital         byte   %8.0g       marital    0=single, 1=married
age             byte   %8.0g                  Mother's age in years
race            byte   %8.0g       race       Race: 1=black, 0=white/other
birthwgt        int    %8.0g                  Birthweight in grams
```

```
. svyset [pweight = finwgt] , strata(stratan)

      pweight: finwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratan
         SU 1: <observations>
        FPC 1: <zero>

. mean birthwgt

Mean estimation                     Number of obs    =    9946

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    birthwgt |   2845.094   9.861422      2825.764    2864.424
--------------------------------------------------------------
```

```
. svy: mean birthwgt
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =       6          Number of obs    =     9946
Number of PSUs   =    9946          Population size  =  3895562
                                    Design df        =     9940

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    birthwgt |   3355.452   6.402741      3342.902    3368.003
--------------------------------------------------------------
```

```
. svyset [pweight = finwgt]

      pweight: finwgt
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>

. svy: mean birthwgt
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =       1          Number of obs    =     9946
Number of PSUs   =    9946          Population size  =  3895562
                                    Design df        =     9945

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    birthwgt |   3355.452   6.933529      3341.861    3369.044
--------------------------------------------------------------

. * So the standard error is, as expected, larger if we ignore the strata.
```

```
. * Now look at multiple regression:

. des race

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
race            byte   %8.0g       race       Race: 1=black, 0=white/other

. gen black = race

. gen married = marital

. tab married

    married |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      4,084       41.03       41.03
          1 |      5,869       58.97      100.00
------------+-----------------------------------
      Total |      9,953      100.00

. gen agesq = age^2

. gen lbirthwgt = log(birthwgt)
(7 missing values generated)
```

```
. svyset [pweight = finwgt], strata(stratan)

      pweight: finwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratan
         SU 1: <observations>
        FPC 1: <zero>

. svy: reg lbirthwgt age agesq black married
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =         6                  Number of obs      =      9946
Number of PSUs     =      9946                  Population size    = 3895561.7
                                                Design df          =      9940
                                                F(   4,   9937)    =    300.19
                                                Prob > F           =    0.0000
                                                R-squared          =    0.0355

------------------------------------------------------------------------------
             |             Linearized
   lbirthwgt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0094712   .0034286     2.76   0.006     .0027504    .0161919
       agesq |  -.0001499   .0000634    -2.36   0.018    -.0002742   -.0000256
       black |   -.074903   .0039448   -18.99   0.000    -.0826356   -.0671703
     married |   .0377781   .0058039     6.51   0.000     .0264013    .0491548
       _cons |   7.941929   .0442775   179.37   0.000     7.855136    8.028722
------------------------------------------------------------------------------
```

```
. svyset [pweight = finwgt]

      pweight: finwgt
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>

. svy: reg lbirthwgt age agesq black married
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =         1                Number of obs      =      9946
Number of PSUs     =      9946                Population size    = 3895561.7
                                              Design df          =      9945
                                              F(   4,   9942)    =    202.34
                                              Prob > F           =    0.0000
                                              R-squared          =    0.0355
------------------------------------------------------------------------------
             |             Linearized
   lbirthwgt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0094712   .0034294     2.76   0.006     .0027489    .0161935
       agesq |  -.0001499   .0000634    -2.36   0.018    -.0002743   -.0000256
       black |   -.074903   .0045443   -16.48   0.000    -.0838106   -.0659953
     married |   .0377781     .00582     6.49   0.000     .0263697    .0491864
       _cons |   7.941929   .0443344   179.14   0.000     7.855024    8.028833
------------------------------------------------------------------------------
```

```
. * Ignore the stratification and VP sampling:

. di .00947/(2*.00015)
31.566667

. reg lbirthwgt age agesq black married, robust

Linear regression                                      Number of obs =    9946
                                                       F(  4,  9941) =   28.56
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0114
                                                       Root MSE      =  .49611

------------------------------------------------------------------------------
             |               Robust
   lbirthwgt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0161755   .0074639     2.17   0.030     .0015448    .0308062
       agesq |  -.0003198    .000138    -2.32   0.020    -.0005902   -.0000493
       black |  -.0136733   .0116097    -1.18   0.239    -.0364307    .0090841
     married |   .0961381   .0129681     7.41   0.000     .0707181    .1215582
       _cons |   7.615568   .0969574    78.55   0.000     7.425512    7.805624
------------------------------------------------------------------------------

. di .0168/(2*.00032)
26.25
```

• Little changes if we allow endogenous explanatory variables in the population model:

$$\begin{aligned} y &= \mathbf{x}\boldsymbol{\beta} + u \\ E(\mathbf{z}'u) &= \mathbf{0} \end{aligned}$$

• From a technical point of view, the estimator is a weighted 2SLS estimator. Or, an optimal weighting matrix can be used in weighted GMM.

• In Stata, the prefix `svy:` supports `ivreg`, too.

## 3. Nonlinear Models

• The same weighting ideas work for a large class of nonlinear models (more precisely, nonlinear estimation methods). In Stata, logit, probit, Tobit, GLM. Currently, not quantile regression.

• In the SS sampling case, the weighted M-estimator solves

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\{\sum_{g=1}^{G}\pi_g\left[N_g^{-1}\sum_{i=1}^{N_g}q(\mathbf{w}_{gi},\boldsymbol{\theta})\right]\right\}. \tag{3.1}$$

- Key representation is

$$E[q(\mathbf{w},\boldsymbol{\theta})] = \pi_1 E[q(\mathbf{w},\boldsymbol{\theta})|\mathbf{w} \in \mathcal{W}_1] + \pi_2 E[q(\mathbf{w},\boldsymbol{\theta})|\mathbf{w} \in \mathcal{W}_2] + \ldots + \pi_G E[q(\mathbf{w},\boldsymbol{\theta})|\mathbf{w} \in \mathcal{W}_G]. \tag{3.2}$$

- In practice, write as

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[ \sum_{i=1}^{N} (\pi_{g_i}/h_{g_i}) q(\mathbf{w}_{gi},\boldsymbol{\theta}) \right] \tag{3.3}$$

where $h_g = N_g/N$. Sometimes the reported weights are scaled differently (without changing the estimation).

• Let $\mathbf{s}(\mathbf{w},\boldsymbol{\theta})$ and $\mathbf{H}(\mathbf{w},\boldsymbol{\theta})$ be the score and Hessian. Asymptotic variance estimator:

$$\left[\sum_{i=1}^{N}(\pi_{g_i}/h_{g_i})\mathbf{H}(\mathbf{w}_{gi},\hat{\boldsymbol{\theta}})\right]^{-1}$$
$$\cdot\left\{\sum_{g=1}^{G}(\pi_g/h_g)^2\left[\sum_{i=1}^{N_g}[\mathbf{s}(\mathbf{w}_{gi},\hat{\boldsymbol{\theta}})-\overline{\hat{\mathbf{s}}_g}][\mathbf{s}(\mathbf{w}_{gi},\hat{\boldsymbol{\theta}})-\overline{\hat{\mathbf{s}}_g}]'\right]\right\} \quad (3.4)$$
$$\cdot\left[\sum_{i=1}^{N}(\pi_{g_i}/h_{g_i})\mathbf{H}(\mathbf{w}_{gi},\hat{\boldsymbol{\theta}})\right]^{-1}.$$

where $\overline{\hat{\mathbf{s}}_g} = N_g^{-1}\sum_{i=1}^{N_g}\mathbf{s}(\mathbf{w}_{gi},\hat{\boldsymbol{\theta}})$ is the within stratum $g$ average of the score.

- In Stata, for many commands, use “svy” prefix after having done “svyset.”

```
svy: logit y x1 ... xK
svy: glm y x1 ... xK, fam(poisson)
```

## 4. General Treatment of Exogenous Stratification

- Suppose our population model specifies some feature of $D(\mathbf{y}|\mathbf{x})$.

Could be a full distribution or a conditional mean.

- To ensure we consistently estimate the solution, say $\boldsymbol{\theta}^*$, to the

population problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} E[q(\mathbf{w}, \boldsymbol{\theta})],$$

we should use weights whether or not stratification is based on

conditioning variables $\mathbf{x}$.

- This is important if we want to recover the "best" approximation to the true underlying model. For example, if we want to estimate $\boldsymbol{\theta}^*$ that provides the best mean square approximation to $E(y|\mathbf{x})$ using misspecified nonlinear least squares, we should use weighting even if it is based only on $\mathbf{x}$.
- Wanting to estimate the solution to the population problem is exactly what is needed for the "double robustness" result for treatment effect estimation using regression plus propensity score weighting.

- However, if we assume that the feature of $D(\mathbf{y}|\mathbf{x})$ is correctly specified, we have chosen an appropriate objective function, and the stratification is based on $\mathbf{x}$, weighting – whether for SS or VP sampling – is not needed for consistency and can be harmful in terms of efficiency.

- For a general treatment, we assume that $\boldsymbol{\theta}_o$ solves

$$\min_{\boldsymbol{\theta} \in \Theta} E[q(\mathbf{w}, \boldsymbol{\theta})|\mathbf{x}], \tag{4.1}$$

for all $\mathbf{x}$. This holds for conditional MLE when the model of the density, $f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$, is correctly specified. It holds for nonlinear least squares with a correctly specified mean. It holds for quasi-MLE in the linear exponential family (LEF) $E(y|\mathbf{x})$ is correctly specified. Also holds for quantile regression when the conditional quantile function is correctly specified.

• The unweighted and weighted estimators are both consistent for $\boldsymbol{\theta}_o$ (for SS and VP sampling).

• Generally, we cannot rank the asymptotic variances of $\hat{\boldsymbol{\theta}}_u$ and $\hat{\boldsymbol{\theta}}_w$. But in one case we can, namely, when the generalized conditional information matrix equality holds: For some $\sigma_o^2 > 0$,

$$E[\nabla_{\boldsymbol{\theta}} q_i(\boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} q_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = \sigma_o^2 E[\nabla_{\boldsymbol{\theta}}^2 q_i(\boldsymbol{\theta}_o)|\mathbf{x}_i]. \tag{4.2}$$

• For (conditional) MLE, $\sigma_o^2 = 1$ [with $q_i(\boldsymbol{\theta}) = -\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})$].

• For NLS, (4.2) holds under $Var(y|\mathbf{x}) = \sigma_o^2$ (homoskedasticity)

• For QMLE in LEF, (4.2) holds under the assumption that the true variance is proportional to the variance in the selected density.

• Without this generalized (conditional) information matrix equality, cannot rank $\hat{\theta}_u$ and $\hat{\theta}_w$. For example, in regression with heteroskedasticity, the weighting for stratification might actually help with heteroskedasticity, too.

• Remember, if weights are used, a “sandwich” covariance matrix is needed even if (4.2) is assumed. (This is one way to see weighting is inefficient.) But with correct specification of the conditional model and exogenous stratification, the strata membership is not needed for asymptotic variance calculation.

$$\widehat{Avar(\hat{\boldsymbol{\theta}}_w)} = \left[ \sum_{i=1}^{N} (\pi_{g_i}/h_{g_i}) \mathbf{H}(\mathbf{w}_{gi}, \hat{\boldsymbol{\theta}}) \right]^{-1}$$

$$\cdot \left\{ \sum_{g=1}^{G} (\pi_g/h_g)^2 \left[ \sum_{i=1}^{N_g} \mathbf{s}(\mathbf{w}_{gi}, \hat{\boldsymbol{\theta}}) \mathbf{s}(\mathbf{w}_{gi}, \hat{\boldsymbol{\theta}})' \right] \right\}$$

$$\cdot \left[ \sum_{i=1}^{N} (\pi_{g_i}/h_{g_i}) \mathbf{H}(\mathbf{w}_{gi}, \hat{\boldsymbol{\theta}}) \right]^{-1}.$$

- The within stratum averages, $\overline{\hat{\mathbf{s}}_g} = N_g^{-1} \sum_{i=1}^{N_g} \mathbf{s}(\mathbf{w}_{gi}, \hat{\boldsymbol{\theta}})$, have disappeared. This is the usual Huber-White sandwich estimator applied to the weighted objective function.